Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 355-356 recibido 1-05-2009; aceptado 5-06-2009

La herramienta IDRA (Indexing and Retrieving Automatically)*

The IDRA (Indexing and Retrieving Automatically) tool

Rubén Granados Muñoz
Facultad de Informática
Univ. Politécnica de Madrid
rgranados@fi.upm.es

Ana García Serrano
ETSI Informática
UNED
agarcia@lsi.uned.es

José M. Goñi Menoyo
ETSI Telecomunicación
Univ. Politécnica de Madrid
josemiguel.goni@upm.es

Resumen: Se presenta brevemente la herramienta IDRA, con licencia GPL 3.0, que a partir de unas funcionalidades básicas, facilita la agregación de nuevas funcionalidades para la investigación en recuperación de información.
Palabras clave: Recuperación de información, motor de búsqueda, visualización de resultados.

Abstract: In this paper the IDRA tool is shortly presented. It has a GPL licence. IDRA Java code facilitates the addition of new functionalities.
Keywords: Information retrieval, search engine, results presentation.

1 Descripción de IDRA

A pesar de la disponibilidad de otros sistemas de indexación (como Lucene), en el grupo de investigación al que pertenecemos, se decidió desarrollar en Java una herramienta abierta a los diferentes formatos y funcionalidades según la aplicación para la que se necesiten realizar tareas relacionadas con la recuperación de información. Además era necesario tener la posibilidad de cambiar los diferentes parámetros y funciones para el cálculo de la similitud o relevancia entre documentos, noticias, informes técnicos o simplemente breves anotaciones de fotos. IDRA ofrece, además de las funcionalidades básicas, otras funcionalidades para gestión y almacenamiento de contenidos de forma eficiente, así como presenta un diseño flexible y bien documentado que facilita su ampliación.

IDRA se distribuye con licencia GPL 3.0 en http://sourceforge.net/projects/idraproject/.

La arquitectura de IDRA está compuesta por los módulos que se describen a continuación.

Módulo de indexación. Realiza la extracción del texto de los documentos, su pre-procesamiento, y la indexación.

La fase de extracción hace uso del software Antiword para los documentos en formato Word, y de la librería de Java PDFBox para aquellos en PDF. El texto estructurado de los ficheros XML se extrae con la librería JDOM.

Entre los pre-procesos aplicados, esta la eliminación de stopwords, signos ortográficos, de puntuación, y caracteres especiales, y el llamado 'preproceso JDOM' para el caso de anotaciones en XML, que evita ciertos errores de JDOM con determinados caracteres.

A partir del texto extraído, se realiza la indexación siguiendo el modelo del espacio vectorial (VSM) [Manning et al 08] utilizando como función de pesado TF-IDF.

Módulo de Recuperación. Recibirá como entrada la consulta del usuario y devolverá una lista con los documentos recuperados, ordenados según la relevancia para la consulta (hasta un valor umbral configurable).

Al texto de la consulta se le aplican los mismos pre-procesos que en la indexación. Luego se construye el vector correspondiente a la consulta, y se calcula la similitud entre dicho

* Parcialmente financiado por el proyecto TIN2007-67407-C03-03 "BRAVO: Búsqueda de Respuestas Avanzada Multimodal y Multilingüe - Recuperación de Información", 2008-10

vector y los vectores de cada uno de los documentos indexados. La similitud entre dos vectores, dentro de la aproximación VSM seguida, es la "cercanía" entre dichos vectores (función coseno).

La herramienta ofrece la posibilidad de lanzar simultáneamente un conjunto de consultas desde un fichero. Se generará un fichero de resultados formado por una línea para cada documento recuperado en la que se indica el número de la consulta, el identificador del documento, y el valor de relevancia (según el formato CLEF, [Granados et al 08]).

Módulo de Gestión del Contenido. Encargado de gestionar la base de datos (BD) de las indexaciones realizadas en el sistema. Permite salvar/recuperar (a/desde fichero) e inicializar la BD en cualquier momento.

Es posible convertir la BD en dos formatos diferentes para facilitar su análisis: (1) *archivo CSV*, para visualizar los datos de la indexación, y (2) *tabla MySQL*, que almacena los datos en los campos de una tabla ya creada en MySQL y que permite realizar consultas SQL.

El contenido de una BD es: (id) identificador del documento; (t) término; (f) nº de apariciones de t en id; (n) frecuencia de t en la colección; (idf) frecuencia inversa de t en la colección; y (w) peso de t en id.

Módulo de Preparación de Datos. Realiza la homogeneización de distintos tipos de datos (colecciones de documentos, lista de consultas de evaluación, lista de respuestas,…) para su posterior tratamiento.

Actualmente, IDRA ofrece la posibilidad de procesar anotaciones XML, como en el caso de la colección de imágenes IAPR TC-12 (http://www.iapr.org/), para ser indexadas seleccionando los campos estructurados que se deseen. También permite realizar las consultas (83) sobre la colección TIME indexada (425 *world news articles from 1963 Time magazine)* (ftp://ftp.cs.cornell.edu/pub/smart/time/).

La herramienta IDRA está abierta para que se incluya cualquier otro tratamiento para estos u otros tipos de datos y su preproceso.

Módulo de Evaluación de Resultados. Facilita el análisis y evaluación de resultados y permite comparar las diferentes configuraciones del sistema o experimentos.

Partiendo del fichero de resultados generado por IDRA para un conjunto de consultas y del de respuestas correctas (qrels), el sistema calcula diferentes parámetros de rendimiento: cobertura o recall, precisión, res01 (secuencia de 0's y 1's que representa los aciertos y los fallos ordenados de la lista de resultados) y AP (*average precision*).

Con estos datos se construyen tablas y/o gráficas para mostrar/comparar resultados.

Módulo Lucene. Indexa y recupera documentos mediante las funcionalidades de la demo del software Lucene. Permite comparar resultados obtenidos por este sistema con los de IDRA.

2 Evaluaciones realizados

La calidad de los resultados obtenidos por IDRA se ha evaluado en diferentes escenarios.

Por un lado, el sistema se presentó en CLEF 2008 [Granados et al 08], en la tarea de recuperación de (20.000) imágenes anotadas, en la que, utilizando únicamente la parte textual, se obtuvo un MAP=0.2253, más alto que el MAP medio obtenido por los 4 mejores experimentos de cada grupo participante.

Por otro lado, el sistema se evaluó utilizando la colección de noticias TIME (425 noticias, 83 consultas). Se obtienen MAP entre 66-74% dependiendo del número de documentos recuperados, mientras que la demo de Lucene está entre 66-73%. Los valores de cobertura y precisión recuperando 20 documentos por consulta también son muy similares para IDRA (90%, 17%) y para Lucene (89%, 17%).

Actualmente se trabaja en el refinamiento de IDRA y la adición de nuevas funcionalidades.

3 Demostración de IDRA

En la demostración se realizarán varios ejercicios y ejecuciones que mostrarán aspectos sobre la robustez y la eficiencia de IDRA (durante un mínimo de 5 minutos) y se guiará y responderá a los asistentes que deseen probar la interfaz y observar los resultados obtenidos.

Bibliografía

Granados R, Benavent X, García-Serrano A, Goñi JM. *MIRACLE-FI at ImageCLEFphoto 2008: Experiences in merging text-based and content-based retrievals*. Working Notes for the CLEF 2008 Workshop.

Manning CD, Raghavan P, Schtze H, *Introduction to information retrieval*, 2008, Cambridge Univ Press New York, NY, USA